Variability in exercise physiology: Can capturing *intra*-individual variation help better understand true *inter*-individual responses?

RUNNING TITLE: *Intra*-individual variation and Inter-individual responses

AUTHORS: Oliver J. Chrzanowski-Smith[1], Eva Piatrikova[1], James A. Betts[1], Sean Williams[1], Javier T. Gonzalez[1]

AFFILIATIONS: [1] Department for Health, University of Bath, Bath, BA2 7AY, UK

CORRESPONDENCE: Dr. Javier T Gonzalez, J.T.Gonzalez@bath.ac.uk

WORD COUNT: 4414

NUMBER OF FIGURES: 2

KEY WORDS: *Intra*-individual variation, *Inter*-individual responses, Variability, Exercise, Training

Abstract

Exploring individual responses to exercise training is a growing area of interest. Understanding reasons behind true observed *inter*-individual responses may help personalise exercise training to maximise the benefits received. While numerous factors have been explored, an often underappreciated consideration in the sport and exercise science field is the influence *intra*-individual variation, both in a single measurement and in response to an intervention, may have on training outcomes. Several study designs and statistical approaches are available to incorporate *intra*-individual variation into interventions and accordingly provide information on whether 'true' *inter*-individual responses are present or if they are an artefact of *intra*-individual variation. However, such approaches are sparingly applied. Moreover, *intra*-individual variation may also be important when true *inter*-individual response differences are present. In this perspective piece, the concept of *intra*-individual variation is described before briefly summarising study designs and statistical practices to account for *intra*-individual variation. We then outline two examples of physiological practices (stratified randomisation and prescribing exercise programmes upon training parameters) to demonstrate why sport and exercise scientists should acknowledge *intra*-individual variation prior to the implementation of an intervention, which potentially offers an additional explanation behind observed true *inter*-individual responses to training. Repeated testing pre-implementation of exercise training would conceptually provide more confident estimates of training parameters, which if utilised in a study design will help attenuate biases that may dictate *inter*-individual differences. Moreover, the incorporation of *intra*-individual differences will facilitate insights into alternative factors that may predict and/or explain *true* observed individual responses to an exercise training programme.

1. Introduction

Observations of *inter*-individual variability and 'non-responders' to physical activity and exercise training have been frequently acknowledged (Mann et al., 2014; Bouchard and Rankinen, 2001). While evidence refuting claims of non-response to both aerobic and resistance exercise exist (Montero and Lundby, 2017; Bonafiglia et al., 2016; Churchward-Venne et al., 2015), interest has grown in attempting to quantify, predict and explain observed *inter*-individual variability in response to interventions (Atkinson, Williamson and Batterham, 2019; Voisin et al., 2018; Sparks, 2017; Hecksteden et al., 2015). Such attempts have involved the application of genomics (Williams et al., 2017; Bouchard et al., 2015), replicated crossover designs (Goltz et al., 2019; Goltz et al., 2018; Senn et al., 2011) and statistical methods (Swinton et al., 2018; Atkinson and Batterham, 2015). Here, we aim to reiterate and demonstrate the importance to sport and exercise scientists in acknowledging *intra*-individual variation.

We first describe the concept of *intra*-individual variation alongside summarising study designs and statistical approaches that incorporate *intra*-individual variation to determine whether true *inter*-individual responses exist. Two examples of common physiological practices are then outlined to illustrate why *intra*-individual variation should be systematically explored prior to the implementation of an exercise training programme. This article extends previous discussions by demonstrating conceptually how *intra*-individual variation in baseline training parameters (peak or maximum oxygen consumption [$\dot{V}O_2$peak and $\dot{V}O_2$max] and lactate threshold) may impact stratified randomisation and the 'exercise dose' prescribed to individuals. To our knowledge, these considerations of *intra*-individual variation have not previously been discussed, yet provide clear and relatable examples of how *intra*-individual variation may contribute to true observed *inter*-individual responses to a training programme.

2. What is *intra*-individual variation?

Intra-individual variation can be defined as the difference in values obtained for an outcome measure(s) when the same participant is studied under similarly standardised testing conditions and procedures. It is also referred to as day-to-day or within-subject variation and provides an indication on the reproducibility or reliability of an observation. Similarly, there is *intra*-individual variation in response to an intervention i.e. variability of pre-to-post differences when the same participant is administered the same intervention. These two types of *intra*-individual variation are inter-connected, derive from three overarching sources and have implications for the design and interpretation of an intervention (see Figure 1).

In practice many physiological observations measured on a continuous scale are composed of a 'true' value plus 'error' (i.e. noise) (Atkinson and Batterham, 2015; Atkinson and Nevill, 1998). This variability, or error, in an estimate can derive from three overarching sources: measurement (or technical) error, biological error and biological variation. Measurement error refers to noise derived from the equipment and protocol used and the experimenter, which theoretically is identical across all individuals (Voisin et al., 2018). Alternatively, biological error derives from the influence of environmental factors such as diurnal variation, sleep quality, diet, or psychological stress (Voisin et al., 2018). Even if such a variable has no measurement error, test-retest variability will likely be prevalent to some extent, attributable to biological noise (Atkinson and Batterham, 2015). Importantly, these 'errors' are distinguishable from biological variation that induces a shift in the true score (e.g. adaptations to training or detraining).

To determine the true *intra*-individual variation of an observation, serial measurements over some time-scale must be conducted (i.e. test-retest, concurrent replicates, day-to-day, trial-to-trial). Repeated measurements within a trial are also necessary if the aim is to distinguish between technical and biological sources of *intra*-individual variation. Similarly, if characterising true *intra*-individual variation in response to an intervention is the aim, then the same intervention must be repeated at least once in the same participants. Repeated measurements will conceptually provide a more accurate estimate of a participant's 'true' value or intervention response, especially when there is no systematic error in measurement (e.g. learning effects or diminishing returns from a training programme). Furthermore, to obtain a more valid measurement of *intra*-individual variation, efforts to reduce all sources of error should be taken, including standardised calibration and testing procedures, appropriate timeframes between testing and adequate pre-trial standardisation on 'determinants' of the outcome variable (e.g. physical activity levels and/or dietary intake). (For detailed discussions on *intra*-individual variation see Swinton et al., 2018; Voisin et al., 2018; Hecksteden et al., 2015; Atkinson and Batterham, 2015; Atkinson and Nevill, 1998). Therefore, to confidently capture *intra*-individual variation many aspects need to be considered.

3. Accounting for *intra*-individual variation to determine whether true *inter*-individual responses to an intervention exist

To infer that true *inter*-individual response differences exist, it is imperative to discern between systematic or 'true' changes (i.e. intervention induced) and *intra*-individual variation (from measurement and biological error) (Solomon, 2018; Voisin et al., 2018). Indeed, *intra*-individual variation is in some circumstances large enough to account for all, or a large proportion of apparent *inter*-individual differences in training responses (e.g. for $\dot{V}O_2$max [Williamson et al., 2017] and weight change [Williamson et al., 2018]). To achieve this distinction several study designs and/or statistical approaches are available that measure *intra*-individual variation and accordingly provide information on whether 'true' *inter*-individual responses are present or if they are an artefact of *intra*-individual variation.

The ideal method is to conduct a replicated randomised controlled trial in the same participants, together with repeated testing within each treatment period (Voisin et al., 2018; Hecksteden et al., 2015; Senn, 2011). Here, participants are randomly allocated to the intervention or control (or the order of receiving these conditions if a crossover design) as per a typical randomised controlled trial (RCT). However, upon completion and after an adequate washout period, the study is essentially repeated in the same participants to examine if individuals demonstrate a consistent response to the intervention relative to control. Clearly this poses considerable logistical and feasibility challenges at both the level of the participant and researcher(s). An alternative is to implement one of these approaches alone i.e. either replicate the intervention or have repeated testing pre- and/or post-trial. While such approaches present similar challenges, several studies have adopted replicated designs (Goltz et al., 2019; 2018; Lindholm et al., 2016; Senn et al., 2011). . For example, Goltz and colleagues (2018) found in a replicated, randomized crossover experimental design that true *inter*-individual differences in subjective appetite and blood hormonal responses to acute exercise were apparent in fifteen healthy males, exceeding measurement error and biological error. Similarly, a more recent randomised replicated cross-over study by Goltz and co-workers (2019) also found true *inter*-individual differences in postprandial appetite responses to a standardised breakfast in eighteen healthy males. Moreover, a similar elegant design was also employed in a knee extension training programme where subjects were their own control through exercising one-leg initially followed by a washout period and then two-leg training (Lindholm et al., 2016). While Lindholm and

co-workers (2016) found the response of a large fraction of genes only changed in one training period, indicating *intra*-individual variation, unfortunately *inter*-individual response differences were not explored. Nevertheless, the appearance of such study designs shows a move towards the importance of measuring *intra*-individual variation to determine whether true *inter*-individual response differences exist.

A further pragmatic compromise is to repeatedly test throughout a trial to act as a surrogate for a repeated intervention (Hecksteden et al., 2018; Hecksteden et al., 2015). Here, serial measurements are ideally obtained at similar intervals throughout an intervention (i.e. a time-series experimental design) where the slope of a linear regression is then fitted to an individual's measured values to determine their response. *Intra*-individual variation can then be calculated as the standard error (i.e. typical error) of an individual's slope in which intervention response (and classification of (non-) responders) can be estimated by pre-determined thresholds (e.g. zero change, or measured day-to-day variability, minimum clinically relevant change or smallest worthwhile difference in the respective outcome variable [Hecksteden et al., 2018; Hecksteden et al., 2015]). This approach can begin to overcome measurement and biological error in the assessment of the intervention response on that occasion but cannot discern how individuals would respond if the intervention were repeated. Furthermore, additional shortcomings to this design exist e.g. the assumption that training adaptations are linear over a programme (Hecksteden et al., 2015), albeit a non-linear regression model (e.g. a mono-exponential curve) can be applied in such circumstances (Bonafiglia et al., 2019), or that the measurement per se does not exhibit a temporal rhythm independent of the intervention. Moreover, Atkinson and colleagues (2019) have recently discussed in-depth several further validity concerns in determining *inter*-individual responses and (non-) responders by counting the number of changes in a sample that exceed or fall below a pre-determined threshold (e.g. sample comparisons of responder counts have low statistical power). Recently, Voisin et al (2018) also highlighted using a control period prior to implementing an intervention. This overcomes potential carry-over effects of exercise training in a repeated intervention and measurements in the control period can act as the baseline. However, treatments are not randomly administered, nor can all sources of variability be disentangled (Voisin et al., 2018).

An overarching shortcoming is also that many of the designs above are not possible for some types of outcome. For example, long-term interventions with "hard" end points (such as RCTs with cardiovascular disease as an end point); or interventions that have learning effects, other similar biases, or require long washout periods. For instance, unaccustomed exercise that elicits marked muscle damage should not be performed as a cross-over, since the repeated bout effect confounds the second-response unless a long washout period is implemented (Goodall et al., 2017; Betts et al., 2009); or similarly, if an intervention supplements lipid soluble antioxidants, many months are required for values to return to un-supplemented levels, by which time the intervention group may no longer be equivalent to the control group. Collectively, this shows that designing an intervention to incorporate *intra*-individual variation involves many complexities.

Alternative statistical approaches can also be applied independently or in adjunct with the above study designs. Atkinson and Batterham (2015) neatly describe how comparing the standard deviation of change between the intervention and control groups can act as a measure of *intra*-individual variation. They demonstrate that *intra*-individual variation can account for a large proportion, if not all, of apparent individual response differences. True individual

responses are only evident, and worth exploring, if the standard deviation for change in the intervention group is substantially larger than the control group.

When a control group is not feasible, a second approach is to calculate the typical error of a measurement (or the within-subject standard deviation) (Solomon, 2018; Swinton et al., 2018). This can be calculated through using difference scores derived from either testing a single participant multiple times or a single test-retest in a group of participants (Swinton et al., 2018). Importantly, repeated testing must occur in a time-frame where the 'true' value should remain theoretically stable (Swinton et al., 2018). Assuming data are normally distributed, the pre-to-post change should be no less than 1.96 standard deviations of the group-level within-subject mean to be 95% confident that the apparent intervention-induced change is not simply *intra*-individual variation (Solomon, 2018). Arguably, alternative reliability statistics could also be used in place of the typical error such as 95% limits of agreement (Bland and Altman, 1986).

Importantly, there are overarching considerations for the above statistical approaches. For example, *intra*-individual variation must be consistent across time (e.g. pre- and post-intervention) and sub-groups / different populations (i.e. display no heteroscedasticity) (Solomon, 2018; Swinton et al., 2018). Similarly, if no true comparator arm is available, standard deviations or typical errors from prior reliability studies can be used (Atkinson and Batterham, 2015), albeit generalisability must then be assumed, which may be troublesome given laboratory specific practices and the often-small sample sizes of such studies (Voisin et al., 2018; Solomon et al., 2018). Moreover, while confounders such as socio-environmental influences, natural variations and certain biases are in principle controlled for by randomisation, it must be assumed no changes in behaviour or other biases have driven any potential pre-to-post differences in the control group. Indeed, controlling for familiarisation effects may pose substantial challenges (e.g. muscle damage induced by unaccustomed exercise [Goodall et al., 2016; Betts et al., 2009]). Furthermore, trial effects (i.e. the Hawthorne effect) can lead to conscious or unconscious changes in behaviour. Such scenarios may skew change scores and misinform interpretations of *intra*-individual variation and subsequently whether true *inter*-individual response differences exist. Nevertheless, the above statistical approaches adjust for error uncertainty in pre-to-post changes, where apparent *inter*-individual response differences are easily able to be encapsulated by *intra*-individual variation.

4. The consideration of *intra*-individual variation prior to a training programme to explain true *inter*-individual responses

The above statistical approaches to quantify *intra*-individual variation employ these methods after data collection. While applying this step is essential to interpret whether further exploration of *inter*-individual responses to an intervention are warranted, if these criteria are met, *intra*-individual variation should not then be neglected. As demonstrated below, *intra*-individual variation should also be considered much earlier in the design and implementation of a training programme as it may be an underlying factor contributing to observed true *inter*-individual responses.

4.1 Example 1: Stratified randomisation

Randomised control trials frequently use stratified randomisation to control for *a priori* identified parameter(s) of importance. This helps reduce confounding influences of co-variates that may mask, attenuate or intensify potential intervention effects and jeopardise conclusions (e.g. regression to the mean or ceiling effects). Consequently, establishing 'true' baseline estimates are imperative (Swinton et al., 2018).

Alongside representing a common outcome measure, cardiorespiratory fitness can be an important baseline characteristic for stratification in an exercise training RCT. Typically, a one-off incremental graded exercise test (GXT) is used to estimate $\dot{V}O_2$peak or $\dot{V}O_2$max as a marker of cardiorespiratory fitness. However, obtaining only a one-off estimate for cardiorespiratory fitness could conceptually threaten stratification. For example, if an individual's estimate of $\dot{V}O_2$peak is assessed only once at baseline, but large variability is unknowingly evident in this estimate, this participant could be categorised into the wrong strata. Repeated assessment at baseline (or the inclusion of a shorter verification protocol [Poole and Jones, 2017]) would in principle provide a more confident estimate of their cardiorespiratory fitness and increase the researcher's confidence that this participant meets the pre-defined strata thresholds. This would consequently attenuate the influence of potential confounding biases (such as selection bias and ceiling effects) that may otherwise be introduced if *intra*-individual variation at baseline was not assessed. In principle this would help to more precisely determine whether true *inter*-individual response differences are apparent and/or facilitate the identification of further contributing factors.

The relevance of this example is apt given findings from studies that have explored the reproducibility of $\dot{V}O_2$peak estimates from GXTs. While high intra-class correlations (0.92–0.99) and low within-subject coefficient of variations (CVs) (3–5%) are typically reported (Edgett et al., 2018; Dideriksen and Mikkelsen, 2017; Midgley et al., 2007), evidence exists that $\dot{V}O_2$peak may be underestimated from an initial or first GXT compared to an identical second and third GXT (Edgett et al., 2018). This learning effect may be particularly evident in individuals inexperienced to maximal testing and importantly influenced the classification of individual responses in $\dot{V}O_2$peak following exercise training (Edgett et al., 2018). Additionally, within-subject CVs and a typical error of up to 9 % and 4.27 $mL \cdot kg^{-1} \cdot min^{-1}$, respectively, for $\dot{V}O_2$max estimates were reported in eleven male amateur runners who completed four identical treadmill GXTs (Lourenço et al., 2011). This demonstrates that *intra*-individual variability in $\dot{V}O_2$peak estimates from one-off GXTs could influence fitness classifications (such as those outlined by Decroix et al. [2016] and De Pauw et al. [2013]). Moreover, a recent study showed group mean estimates of $\dot{V}O_2$peak varied by ~1–5 $mL \cdot kg^{-1} \cdot min^{-1}$ alongside within-subject CVs between 2.0 – 5.2 %, when five different GXT protocols employing varying stage lengths were compared in seventeen trained male cyclists (Jamnick et al., 2018).

To further demonstrate the potential impact that *intra*-individual variation in a baseline characteristic may have for stratified randomisation, a theoretical example is provided in Figure 2. This figure reflects a hypothetical scenario where participants $\dot{V}O_2$peak ($mL \cdot kg^{-1} \cdot min^{-1}$) has been estimated at baseline on three separate occasions (GXT 1, 2 and 3) from the same treadmill GXT. The within-subject variability of $\dot{V}O_2$peak is within the typical error reported by Lourenco and colleagues (2011) i.e. 4.27 $mL \cdot kg^{-1} \cdot min^{-1}$, where stratified randomisation is to be performed for participants who have a $\dot{V}O_2$peak threshold of < 45 $mL \cdot kg^{-1} \cdot min^{-1}$ (threshold derived from performance level 1 fitness classification in males as outlined by De Pauw et al. (2013)). As illustrated, for participant's 1, 4, 5, 6 and 10, if $\dot{V}O_2$peak was assessed only once at baseline (i.e. GXT 1), the researcher(s) would assume these participants are similarly matched for cardiorespiratory fitness and would believe stratified randomisation, to say an exercise RCT, is appropriate. However, if *intra*-individual variability was accounted for by repeated assessment at baseline (i.e. obtaining an average from each individual's GXT 1, GXT 2 and GXT 3 values), a more precise estimate of the participant's true fitness levels (e.g. the within-subject mean on Figure 2) would conceptually be obtained. The researcher(s) would then see that they would be incorrect to perform stratified randomisation on participant 1, 4

and 10. Equally, the reverse is true for participant 2 and 8, who initially would be excluded from stratified randomisation based on the observed value from GXT 1, but in actual fact could be appropriately stratified were repeated assessment to be performed. While the meaningfulness of ± 4.27 $mL \cdot kg^{-1} \cdot min^{-1}$ in $\dot{V}O_2peak$ could be questioned, the relevance of this variability is highlighted by a meta-analysis of n = 34 studies that reported sprint interval training (mean intervention length of 5-weeks) improved $\dot{V}O_2peak$ by 8 % (Vollaard et al., 2017), which equates to 3.5 $mL \cdot kg^{-1} \cdot min^{-1}$ with the $\dot{V}O_2peak$ threshold used above. This hypothetical example shows how overlooking *intra*-individual variation in a baseline characteristic could in principle lead to inappropriate stratified randomisation and introduce biases that may affect analysis techniques (e.g. skew the standard deviation of change in the intervention and/or control groups) and mask, attenuate or intensify intervention effects and *inter*-individual response differences to exercise training.

Collectively, this suggests repeated assessment (or verification tests) are necessary to obtain more confident estimates of baseline characteristics to stratify upon. Moreover, given the potential influence of learning effects, researchers and practitioners may wish to determine the number of assessments required for this bias to dissipate (and consequently exclude initial measurements as appropriate) and/or then obtain the average of the remaining repeated measurements. This arguably would facilitate a more confident assessment of baseline parameters, where acknowledging *intra*-individual variation prior to randomisation may assist with participant group allocation and consequently help remove further confounding biases that may contribute to observed true *inter*-individual responses.

4.2 Example 2: Standardisation of prescribed exercise dose

Many exercise training programmes and RCTs 'standardise' the exercise dose i.e. the workload performed by participants, by fixing the exercise intensity, duration and/or frequency of sessions between participants. However, the method used to standardise exercise programmes varies considerably, leading to concerns over whether the exercise dose standardisation procedure allows precise quantification of *inter*-individual responses (Ross et al., 2019). In a similar manner, *intra*-individual variation in training prescription parameters may pose a concern not only for the standardisation of exercise dose between-subjects but also within a participant during an exercise programme. To our knowledge, the potential implication of *intra*-individual variation in training parameters that are used to prescribe exercise dose has not previously been highlighted but may contribute to observed true *inter*-individual response differences.

To demonstrate the importance of acknowledging *intra*-individual variation in training prescription parameters, a hypothetical example is provided whereby a training programme prescribes participants a *'set'* relative intensity to exercise at derived from a one-off GXT. As issues of prescribing exercise intensity based on a percentage of $\dot{V}O_2max$, or a percentage / beats below maximum heart rate (HR_{MAX}) have been discussed elsewhere (Piatrikova et al., 2019; Mann et al., 2013; Meyer et al., 1999), this example focuses on the recommendation to prescribe exercise upon indices that elicit more similar physiological responses between-subjects such as the lactate threshold or critical speed.

Before describing this scenario, it is important to acknowledge that the precise prescription of exercise intensity is particularly important given that the physiological responses to exercise intensity are not necessarily linear. If the physiological stress displayed a linear relationship across all exercise intensities, then (non-systematic) variability could be reduced simply with randomisation and a sufficient sample size. However, since the metabolic stress response to

exercise is non-linear, an over-estimation of exercise intensity could disproportionally affect the physiological response compared to an equivalent under-estimate, and therefore balance would not necessarily be achieved by randomisation. Accordingly, repeated assessment at baseline to accurately prescribe exercise intensity (and at time-points throughout a training programme to recalibrate the prescribed exercise intensity to account for any training adaptations) can be important to ensure that the adaptive stimuli is similar across people within each group of an intervention.

Take a hypothetical situation where *intra*-individual variability in the GXT used to determine the lactate threshold, for which exercise training sessions are prescribed upon, is unknowingly large. The metabolic stress (i.e. 'training stimuli') induced by each acute exercise bout may consequently vary session-to-session. In support of this example, the corresponding speed and heart rate at which the lactate threshold (first significant elevation of blood lactate concentration above resting levels) and fixed 4 $mmol \cdot L^{-1}$ blood lactate concentration were detected, showed 95% limits of agreement of +/-± 1.5 and 1.3 $km \cdot h^{-1}$ and 16 and 12 beats per minute, respectively in twenty males and sixteen females who were young, healthy and active (Grant et al., 2002). This variability in running speed at "lactate threshold" is equivalent to ~10%, which is therefore substantial. Similar low reproducibility in several blood lactate markers during GXTs have also subsequently been reported, albeit partly moderated by factors such as analysis method, stage duration and training status (Gavin et al., 2014; Morton et al., 2012). Training status is particularly important given that sedentary individuals are often recruited to training programmes, where reproducibility of lactate measures are speculated to be lower (Gavin et al., 2014; Grant et al., 2002). Further support for the realism of the above example derives from a recent study that found substantial *inter*-method variability when estimating the lactate threshold via five one-off GXT protocols of various stage lengths and fourteen analysis techniques in seventeen trained males (Jamnick et al., 2018).

Echoing the issue of prescribing relative exercise intensity upon $\dot{V}O_2max$ or HR, the potential variability in 'training stimuli' session-to-session may induce different training adaptations, supporting previous speculations and potentially accounting for observations of 'responders' and 'non-responders' to a training programme (Mann et al., 2013; 2014). Moreover, this potential variability in training stimuli may influence the standard deviation of change in the intervention group and have important implications for data interpretation (Voisin et al., 2018). Further complications may also derive from individuals potentially having different capacities to work aerobically and anaerobically (Piatrikova et al., 2018; Buchheit and Laursen, 2013). The applicability of this example is apt given preliminary findings that acute differences in metabolic stress to the first exercise training session (mean blood lactate concentrations) were positively associated (via a simple linear regression) with increases in $\dot{V}O_2peak$ after 4-weeks of exercise training (Preobrazenski et al., 2018), albeit approaches to adjust for *intra*-individual variation in pre-to-post changes were not employed. Nevertheless, the above collectively suggests that a more confident estimate of the selected parameter to prescribe training upon would conceptually provide more assurance that participants are exercising at an intensity that elicits similar physiological responses both within- and between-subjects. This can be achieved by repeated testing prior to the implementation of and during a training programme, which would arguably lead to a more precise standardisation of the exercise dose prescribed. Collectively, this would attenuate any potential confounding bias introduced by *intra*-individual variation that may contribute to true observed *inter*-individual responses to a training programme.

5. Conclusion

This perspective piece highlights the importance that *intra*-individual variation in baseline and training parameters may have on the implementation of a training programme and consequently, how this may dictate apparent group and true *inter*-individual responses to a training programme. Ultimately, the reasons behind *true* heterogeneous training adaptations are likely multi-dimensional (Solomon, 2018; Swinton et al., 2018; Hecksteden et al., 2015) and there is unlikely one universal solution to incorporate *intra*-individual variation (Hecksteden et al., 2015). Nevertheless, while quantifying and controlling for *intra*-individual variation through repeated testing is undoubtedly challenging, researchers who do this will be better placed to: a) identify *true* effects of a training programme and b) more confidently and appropriately prescribe 'personalised' training programmes on an individual basis. Moreover, while examples specific to aerobic endurance training were used, the implications of *intra*-individual variation highlighted here are highly applicable and transferable to all domains of sport and exercise science (e.g. resistance exercise, biomechanics and / or psychology). Overall, acknowledging *intra*-individual variation will attenuate a potential confounding variable and facilitate greater insights into alternative variables that may predict and/or explain true observed *inter*-individual responses to exercise training.

Figures

Figure 1. Sources and potential implications of *intra*-individual variation

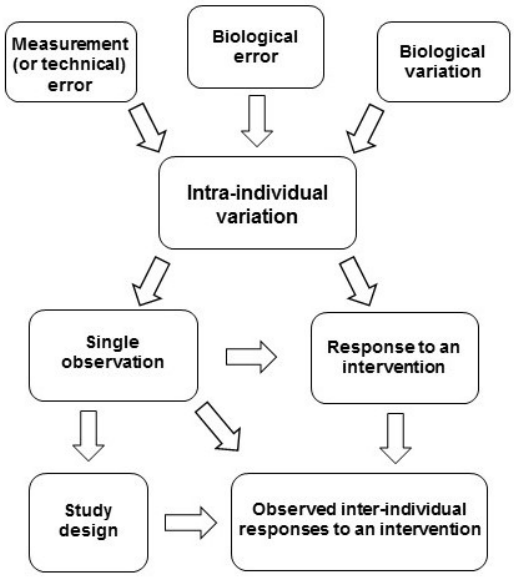
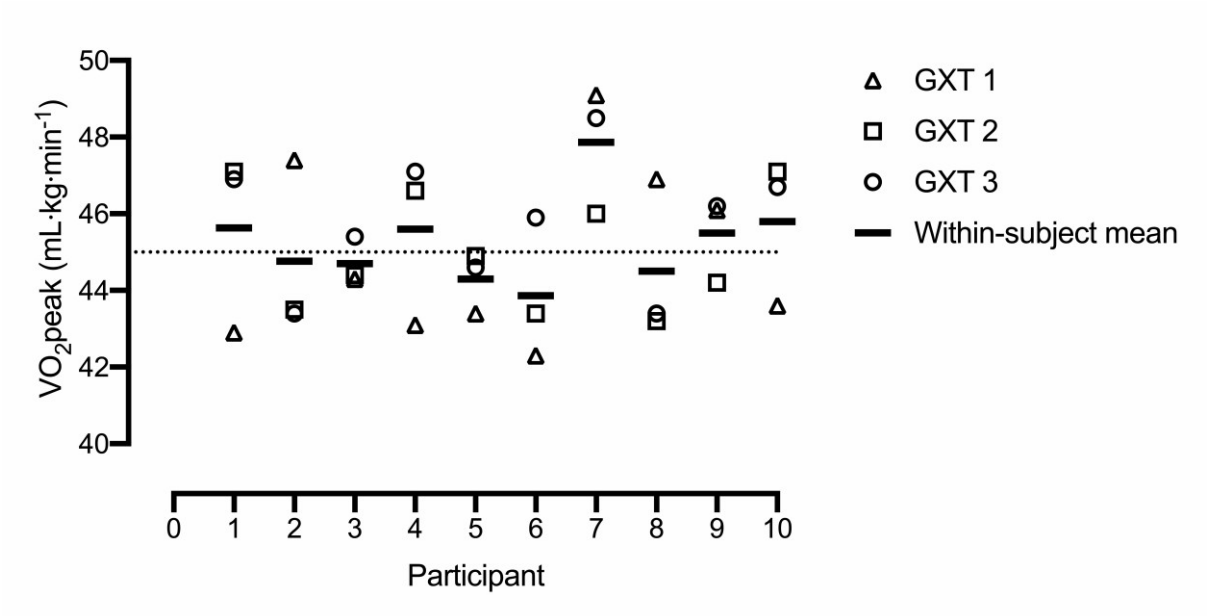


Figure 2. A hypothetical scenario to demonstrate the influence *intra*-individual variation at baseline may have for stratified randomisation



Author Contributions

The manuscript was written by O.C-S. All authors (E.P., J.B., S.W. and J.G.) contributed to each section and revised and approved the final manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Declarations of Funding Sources

Funding was provided by the Department for Health and University Research Studentship from the Graduate School, Faculty of Humanities and Social Sciences at the University of Bath. Funds were not involved in any part of writing the manuscript except for providing office equipment and financial support.

References

Atkinson, G., Williamson, P., & Batterham, A. M. (2019). Issues in the determination of 'responders' and 'non-responders' in physiological research. Exp. Physiol. doi:10.1113/ep087712

Atkinson, G., & Batterham, A. M. (2015). True and false interindividual differences in the physiological response to an intervention. Exp. Physiol., 100(6), 577-588. doi:10.1113/ep085070

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med., 26(4), 217-238.

Bonafiglia, J. T., Ross, R., & Gurd, B. J. (2019). The application of repeated testing and monoexponential regressions to classify individual cardiorespiratory fitness responses to exercise training. Eur. J. Appl. Physiol. 119(4), 889-900. doi:10.1007/s00421-019-04078-w

Betts, J. A., Toone, R. J., Stokes, K. A., & Thompson, D. (2009). Systemic indices of skeletal muscle damage and recovery of muscle function after exercise: effect of combined carbohydrate-protein ingestion. Appl. Physiol. Nutr. Metab., 34(4), 773-784. doi:10.1139/h09-070

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1(8476), 307-310.

Bonafiglia, J. T., Rotundo, M. P., Whittall, J. P., Scribbans, T. D., Graham, R. B., & Gurd, B. J. (2016). Inter-Individual Variability in the Adaptive Responses to Endurance and Sprint Interval Training: A Randomized Crossover Study. PloS one, 11(12), e0167790. doi:10.1371/journal.pone.0167790

Bouchard, C., Antunes-Correa, L. M., Ashley, E. A., Franklin, N., Hwang, P. M., Mattsson, C. et al. (2015). Personalized preventive medicine: genetics and the response to regular exercise in preventive interventions. Prog. Cardiovasc. Dis., 57(4), 337-346. doi:10.1016/j.pcad.2014.08.005

Bouchard, C., & Rankinen, T. (2001). Individual differences in response to regular physical activity. Med. Sci. Sports Exerc., 33(6 Suppl), S446-451; discussion S452-443.

Buchheit, M., & Laursen, P. B. (2013). High-intensity interval training, solutions to the programming puzzle: Part I: cardiopulmonary emphasis. Sports Med., 43(5), 313-338. doi:10.1007/s40279-013-0029-x

Churchward-Venne, T. A., Tieland, M., Verdijk, L. B., Leenders, M., Dirks, M. L., de Groot, L. C., et al. (2015). There Are No Nonresponders to Resistance-Type Exercise Training in Older Men and Women. J. Am. Med. Dir. Assoc., 16(5), 400-411. doi:10.1016/j.jamda.2015.01.071

De Pauw, K., Roelands, B., Cheung, S. S., de Geus, B., Rietjens, G., & Meeusen, R. (2013). Guidelines to classify subject groups in sport-science research. Int. J. Sports. Physiol. Perform., 8(2), 111-122.

Decroix, L., De Pauw, K., Foster, C., & Meeusen, R. (2016). Guidelines to Classify Female Subject Groups in Sport-Science Research. Int. J. Sports. Physiol. Perform., 11(2), 204-213. doi:10.1123/ijspp.2015-0153

Dideriksen, K., & Mikkelsen, U. R. (2017). Reproducibility of incremental maximal cycle ergometer tests in healthy recreationally active subjects. Clin. Physiol. Funct. Imaging., 37(2), 173-182. doi:10.1111/cpf.12283

Edgett, B. A., Bonafiglia, J. T., Raleigh, J. P., Rotundo, M. P., Giles, M. D., Whittall, J. P., et al. (2018). Reproducibility of peak oxygen consumption and the impact of test variability on classification of individual training responses in young recreationally active adults. Clin. Physiol. Funct. Imaging, 38(4), 630-638. doi:10.1111/cpf.12459

Gavin, J. P., Willems, M. E., & Myers, S. D. (2014). Reproducibility of lactate markers during 4 and 8 min stage incremental running: a pilot study. J. Sci. Med. Sport, 17(6), 635-639. doi:10.1016/j.jsams.2013.08.006

Goltz, F. R., Thackray, A. E., Atkinson, G., Lolli, L., King, J. A., Dorling, J. L., Dowejko, M., Mastana, S., & Stensel, D. J. (2019). True Interindividual Variability Exists in Postprandial Appetite Responses in Healthy Men But Is Not Moderated by the FTO Genotype. J. Nutr. doi:10.1093/jn/nxz062

Goltz, F. R., Thackray, A. E., King, J. A., Dorling, J. L., Atkinson, G., & Stensel, D. J. (2018). Interindividual Responses of Appetite to Acute Exercise: A Replicated Crossover Study. Med. Sci. Sports Exerc., 50(4), 758-768. doi:10.1249/mss.0000000000001504

Goodall, S., Thomas, K., Barwood, M., Keane, K., Gonzalez, J. T., St Clair Gibson, A., et al. (2017). Neuromuscular changes and the rapid adaptation following a bout of damaging eccentric exercise. Acta Physiol. (Oxf.), 220(4), 486-500. doi:10.1111/apha.12844

Grant, S., McMillan, K., Newell, J., Wood, L., Keatley, S., Simpson, D., et al., (2002). Reproducibility of the blood lactate threshold, 4 mmol.l(-1) marker, heart rate and ratings of perceived exertion during incremental treadmill exercise in humans. Eur. J. Appl. Physiol., 87(2), 159-166. doi:10.1007/s00421-002-0608-2

Hecksteden, A., Kraushaar, J., Scharhag-Rosenberger, F., Theisen, D., Senn, S., & Meyer, T. (2015). Individual response to exercise training - a statistical perspective. J. Appl. Physiol. (1985), 118(12), 1450-1459. doi:10.1152/japplphysiol.00714.2014

Hecksteden, A., Pitsch, W., Rosenberger, F., & Meyer, T. (2018). Repeated testing for the assessment of individual response to exercise training. J. Appl. Physiol. (1985), 124(6), 1567-1579. doi:10.1152/japplphysiol.00896.2017

Jamnick, N. A., Botella, J., Pyne, D. B., & Bishop, D. J. (2018). Manipulating graded exercise test variables affects the validity of the lactate threshold and [Formula: see text]. PloS one, 13(7), e0199794. doi:10.1371/journal.pone.0199794

Lindholm, M. E., Giacomello, S., Werne Solnestam, B., Fischer, H., Huss, M., Kjellqvist, S., et al. (2016). The Impact of Endurance Training on Human Skeletal Muscle Memory, Global Isoform Expression and Novel Transcripts. PLoS Genet., 12(9), e1006294. doi:10.1371/journal.pgen.1006294

Lourenco, T. F., Martins, L. E., Tessutti, L. S., Brenzikofer, R., & Macedo, D. V. (2011). Reproducibility of an incremental treadmill VO(2)max test with gas exchange analysis for runners. J Strength Cond Res, 25(7), 1994-1999. doi:10.1519/JSC.0b013e3181e501d6

Mann, T., Lamberts, R. P., & Lambert, M. I. (2013). Methods of prescribing relative exercise intensity: physiological and practical considerations. Sports Med., 43(7), 613-625. doi:10.1007/s40279-013-0045-x

Mann, T. N., Lamberts, R. P., & Lambert, M. I. (2014). High responders and low responders: factors associated with individual variation in response to standardized training. Sports. Med., 44(8), 1113-1124. doi:10.1007/s40279-014-0197-3

Meyer, T., Gabriel, H. H., & Kindermann, W. (1999). Is determination of exercise intensities as percentages of VO2max or HRmax adequate? Med. Sci. Sports. Exerc., 31(9), 1342-1345.

Midgley, A. W., McNaughton, L. R., & Carroll, S. (2007). Effect of the VO2 time-averaging interval on the reproducibility of VO2max in healthy athletic subjects. Clin Physiol Funct Imaging, 27(2), 122-125. doi:10.1111/j.1475-097X.2007.00725.x

Montero, D., & Lundby, C. (2017). Refuting the myth of non-response to exercise training: 'non-responders' do respond to higher dose of training. J. Physiol., 595(11), 3377-3387. doi:10.1113/jp273480

554 Morton, R. H., Stannard, S. R., & Kay, B. (2012). Low reproducibility of many lactate markers
555 during incremental cycle exercise. Br. J. Sports Med., 46(1), 64-69.
556 doi:10.1136/bjsm.2010.076380
557 Piatrikova, E., Sousa, A., Willsmer, N., Gonzalez, J. T., & Williams, S. (2019). Demarcating
558 exercise intensity domains in freestyle swimming; is there an alternative to incremental
559 step test and beats below HRMAX method? J. Strength. Cond. Res. doi:
560 10.1519/JSC.0000000000002884
561 Piatrikova, E., Sousa, A. C., Gonzalez, J. T., & Williams, S. (2018). Validity and Reliability of
562 the 3-Minute All-Out Test in National and International Competitive Swimmers. Int. J.
563 Sports. Physiol. Perform., 1-24. doi:10.1123/ijspp.2018-0018
564 Poole, D. C., & Jones, A. M. (2017). Measurement of the maximum oxygen uptake Vo2max:
565 Vo2peak is no longer acceptable. J. Appl. Physiol. (1985), 122(4), 997-1002.
566 doi:10.1152/japplphysiol.01063.2016
567 Preobrazenski, N., Bonafiglia, J. T., Nelms, M. W., Lu, S., Robins, L., LeBlanc, C., et al.,
568 (2018). Does blood lactate predict the chronic adaptive response to training: A
569 comparison of traditional and talk test prescription methods. Appl. Physiol. Nutr. Metab.
570 doi:10.1139/apnm-2018-0343
571 Ross, R., Goodpaster, B. H., Koch, L. G., Sarzynski, M. A., Kohrt, W. M., Johannsen, N. M.,
572 Skinner, J. S., Castro, A., Irving, B. A., Noland, R. C., Sparks, L. M., Spielmann, G.,
573 Day, A. G., Pitsch, W., Hopkins, W.G., & Bouchard, C. (2019). Precision exercise
574 medicine: understanding exercise response variability. Br. J. Sports Med.
575 doi:10.1136/bjsports-2018-100328
576 Senn, S., Rolfe, K., & Julious, S. A. (2011). Investigating variability in patient response to
577 treatment--a case study from a replicate cross-over study. Stat Methods Med Res, 20(6),
578 657-666. doi:10.1177/0962280210379174
579 Solomon, T. P. J. (2018). Sources of Inter-individual Variability in the Therapeutic Response
580 of Blood Glucose Control to Exercise in Type 2 Diabetes: Going Beyond Exercise Dose.
581 Front Physiol, 9, 896. doi:10.3389/fphys.2018.00896
582 Sparks, L. M. (2017). Exercise training response heterogeneity: physiological and molecular
583 insights. Diabetologia, 60(12), 2329-2336. doi:10.1007/s00125-017-4461-6
584 Swinton, P. A., Hemingway, B. S., Saunders, B., Gualano, B., & Dolan, E. (2018). A Statistical
585 Framework to Interpret Individual Response to Intervention: Paving the Way for
586 Personalized Nutrition and Exercise Prescription. Front Nutr, 5, 41.
587 doi:10.3389/fnut.2018.00041
588 Vollaard, N. B. J., Metcalfe, R. S., & Williams, S. (2017). Effect of Number of Sprints in an
589 SIT Session on Change in V O2max: A Meta-analysis. Med. Sci. Sports Exerc., 49(6),
590 1147-1156. doi:10.1249/mss.0000000000001204
591 Voisin, S., Jacques, M., Lucia, A., Bishop, D. J., & Eynon, N. (2018). Statistical Considerations
592 for Exercise Protocols Aimed at Measuring Trainability. Exerc. Sport. Sci. Rev.
593 doi:10.1249/jes.0000000000000176
594 Williams, C. J., Williams, M. G., Eynon, N., Ashton, K. J., Little, J. P., Wisloff, U., et al.,
595 (2017). Genes to predict VO2max trainability: a systematic review. BMC Genomics,
596 18(Suppl 8), 831. doi:10.1186/s12864-017-4192-6
597 Williamson, P. J., Atkinson, G., & Batterham, A. M. (2017). Inter-Individual Responses of
598 Maximal Oxygen Uptake to Exercise Training: A Critical Review. Sports Med., 47(8),
599 1501-1513. doi:10.1007/s40279-017-0680-8
600 Williamson, P. J., Atkinson, G., & Batterham, A. M. (2018). Inter-individual differences in
601 weight change following exercise interventions: a systematic review and meta-analysis
602 of randomized controlled trials. Obes. Rev., 19(7), 960-975. doi:10.1111/obr.12682